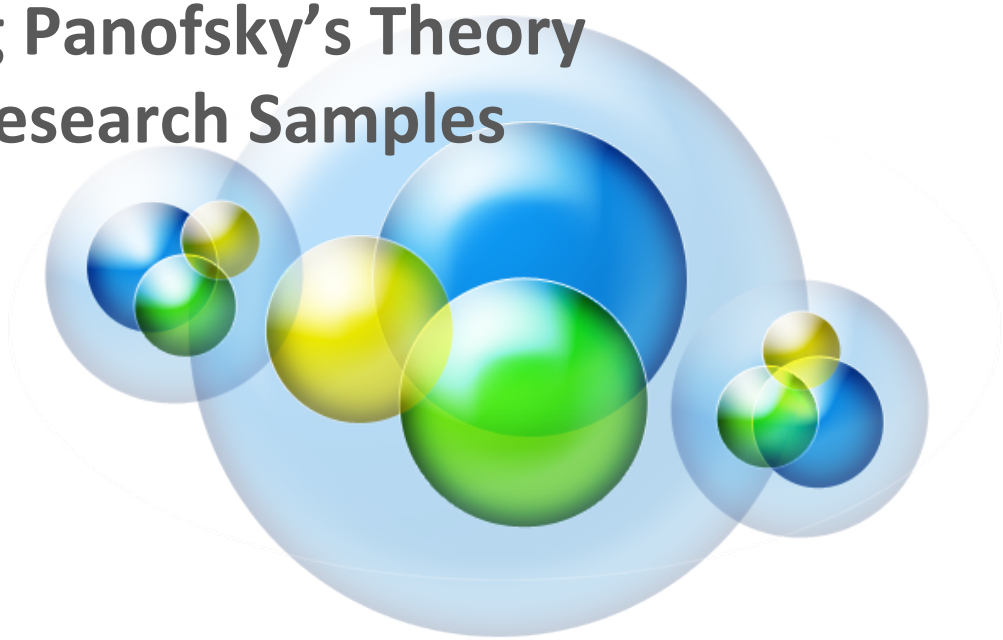
Using a Semantic Analysis Tool to Generate Subject Access Points:

A Study using Panofsky's Theory and Two Research Samples



Marcia Lei Zeng

Karen F. Gracy,

Kent State University, USA

&

Maja Žumer

University of Ljubljana, Slovenia

13th International Conference (ISKO 2014)
Krakow, Poland,. May 19th-22nd 2014

Background

The Research Question:

- the assessment of alternative approaches of generating subject access points to the materials that are usually not made available through regular library catalog routines.
  - Subject access is critical for cross-institutional digital libraries.
  - Limited subject access points are particularly critical with very large-scale resources of cross-institutional collections.
  - LAMs are recognizing the impracticality and impossibility of conducting exhaustive traditional subject analysis.

Example of alternative:
-- Natural language processing and semantic annotation

From an original surrogate:

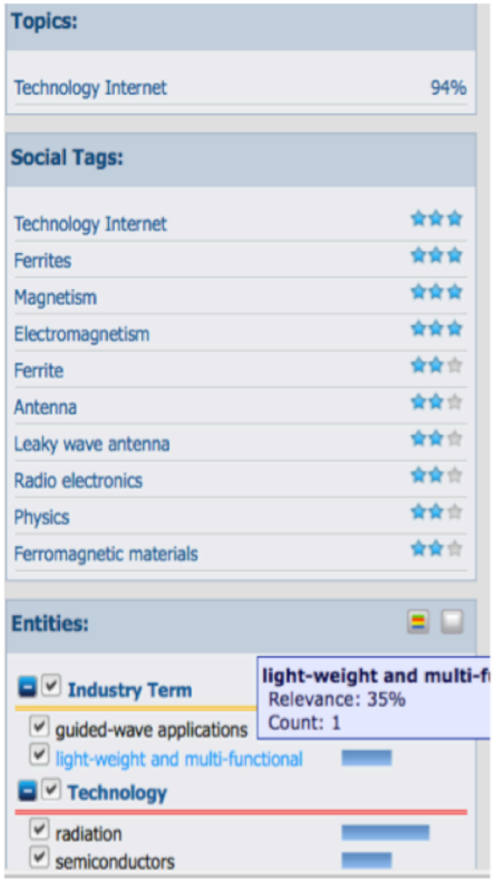
Title: Novel Implementations of Coupled Microstrip Lines on Magnetic Substrates

Pages
216 p.

Subject Headings
Electrical Engineering; Electromagnetism

Keywords
leaky-wave antennas, ferrite, magnetic materials, anisotropy, beamsteering, magnetic beamsteering, nonreciprocity

After running through the tool:



software-suggested access points:

- named entities
- topics)
- relations of the contents of a given resource

Panofsky's three-layer framework and the simplified layers used by CCO

Object of Interpretation	Act of Interpretation	Equipment for Interpretation	Controlling principle of Interpretation	Simplified layers [2]
I-*Primary or natural* subject matter – (A) factual, (B) expressional-, constituting the world of artistic motifs	*Pre-iconographical description* (and pseudo-formal analysis).	*Practical experience* (familiar with *objects* and *events*).	History of *style* (insight into the manner in which, under varying historical conditions, *objects* and *events* were expressed by *forms*).	I-Description (refer to the generic elements depicted in or by the work).
II-*Secondary* or *conventional* subject matter, constituting the world of *images*, *stories* and *allegories*.	*Iconographical analysis* in the narrower sense of the word.	*Knowledge* of *literary sources* (familiar with specific *themes* and *concepts*).	History of *types* (insight into the manner in which, under varying historical conditions, specific *themes* or *concepts* were expressed by *objects* and *events*).	II-Identification (refer to the specific subject).
III-*Intrinsic meaning* or *content*, constituting the world of '*symbolical' values*.	*Iconographical interpretation* in a deeper sense (*iconographical synthesis*)	*Synthetic intuition* (familiar with the *essential tendencies* of *the human mind*), conditioned by personal psychology and '*Weltanschauung*.'	History of *cultural symptoms* or '*symbols*' in general (insight into the manner in which, under varying historical conditions, *essential tendencies of –the- human mind* were expressed by specific *themes* and *concepts*).	III – Interpretation (refer to the meaning or themes represented by the subjects and includes a conceptual analysis of what the work is about).

An example of

Using a Semantic Analysis Tool to Generate Subject Access Points

Finding Aids : Pearl Harbor Attack (Dec 6-Dec 8, 1941)

Page: 1 of 6 140%

Finding Aid to the Artificial Collection:

Pearl Harbor Attack (Dec 6 – Dec 8, 1941)

Size: (.5 cu.ft.)

Dates: December 6, 1941 – December 8, 1941

Location of Repository: Franklin D. Roosevelt Presidential Library

Name of Finding Aid Author: Ali Caron & Georgina Garcia

Date of Creation: Summer 2011

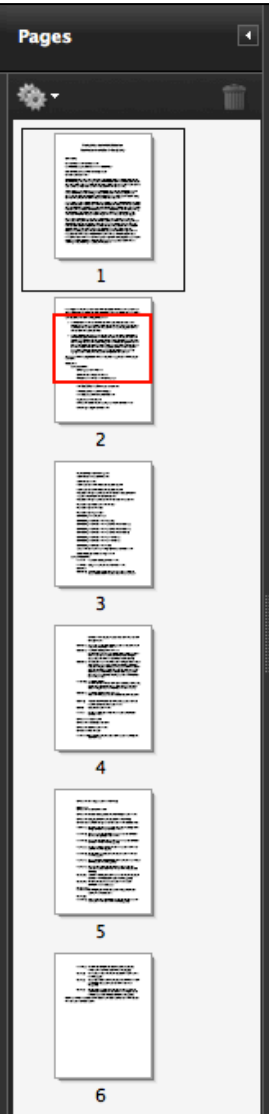
Copyright Notice: The writings of Franklin D. Roosevelt within this collection are in the public domain. The official writings of United States government officials within this collection are in the public domain. The writings of Eleanor Roosevelt within this collection are subject to Mrs. Roosevelt's literary estate. All other materials are subject to the United States Copyright law, 17 U.S.C. 101, et seq.

Administrative Note: Franklin D. Roosevelt Presidential Library is the first presidential library and the only presidential library used by a sitting president. The library houses documents, objects, and photographs pertaining and relating to President Franklin D. Roosevelt (FDR). FDR was in office during the events of Pearl Harbor and the library houses materials about the attack.

Scope and Content: This artificial collection is composed of photocopies of original documents identified

Source: http://www.fdrlibrary.marist.edu/archives/pdfs/findingaids/findingaid_pearlharborattack.pdf

(cont.) Finding Aids : Pearl Harbor Attack (Dec 6-Dec 8, 1941)



Series Descriptions: The collection is organized in 2 series:

- Series I: Documents – The items selected for this series remain within December 6, 1941- December 8, 1941, date range. The items reflect the Pearl Harbor attack or events relevant to that incident. The contents found under the container list portion are arranged by: collection title; and folder title, found in quotations.

- Series II: Still Photographs – Images comprising this series are selected from the FDR Library's General Photograph Collection, folder: WWII: Hawaii: Attack on Pearl Harbor. Listed here are original captions taken from the photographs themselves, along with a Library control number. Unless copyright information is stated in the image caption, all of the photographs in this series belong in the public domain. This means that, to the best of our knowledge, the materials may be freely used by the researcher. However, for copyrighted materials, it is the researcher's responsibility to determine the limits of Fair Use as defined by sections 107 to 118 of the copyright law and to obtain permission from the copyright holder for further use.

Arrangement: Series I is arranged alphabetically after the president's papers and Series II is arranged numerically.

Container Lists:

SERIES I: DOCUMENTS

OF400: Appointments; Hawaii, 1941

OF4675: World War II; General, 1941-1942

PPF200b: Nov. 11, 1941- Jan. 6, 1942; Public Reactions

Source: http://www.fdrlibrary.marist.edu/archives/pdfs/findingaids/findingaid_pearlharborattack.pdf

CALAIS
Powered by Thomson Reuters

The Calais initiative is about enabling semantic applications by providing a metadata generation web service, sample applications using that service to jumpstart development efforts, and support for

Enter text here:

```
Finding Aid t  the Artificial Collection:
Pearl Harbor Attack (Dec 6 – Dec 8, 1941)
Size: (.5 cu.ft.)
Dates: December 6, 1941 – December 8, 1941
Location of Repository: Franklin D. Roosevelt Presidential Library
Name of Finding Aid Author: Ali Caron & Georgina Garcia
Date of Creation: Summer 2011
Copyright Notice: The writings of Franklin D. Roosevelt within this collection are in the
public domain. The official writings of United States government officials within this
collection are in the public domain. The writings of Eleanor Roosevelt within this
collection are subject to Mrs. Roosevelt's literary estate. All other materials are
```

Social Tags:

- Hawaii
- Film
- Attack on Pearl Harbor
- Pearl Harbor advance-knowledge conspiracy theory
- National Pearl Harbor Remembrance Day
- USS Arizona Memorial
- Battleship Row
- Pearl Harbor
- Geography of the United States

Entities:

- City
- Company
- Country
- Date
- Facility

Finding Aid to the Artificial Collection:

Pearl Harbor Attack (Dec 6 – Dec 8, 1941)

Size: (.5 cu.ft.)

Dates: December 6, 1941 – December 8, 1941

Location of Repository: **Franklin D. Roosevelt** Presidential Library

Name of **Finding Aid Author**: **Ali Caron** & **Georgina Garcia**

Date of Creation: Summer 2011

Copyright Notice: The writings of **Franklin D. Roosevelt** within this collection are in the public domain. **The official writings** of **United Sta**

e public domain. The writings of **Eleanor Roosevelt** within this collection are subject to **Mrs. Roosevelt**'s literary estate. All other materials

101, et seq.

Administrative Note: **Franklin D. Roosevelt Presidential Library** is the first presidential library and the only presidential library used by **a s**

, and photographs pertaining and relating to **President Franklin D. Roosevelt** (FDR). FDR was in office during the events of **Pearl Harbor**

Scope and Content: This artificial collection is composed of photocopies of original documents identified from the holdings of the Franklin D. R

specifically the events of three days: December 6, 7, and 8, 1941. The items are a collection of documents gathered from other collections fou

The criterion for selecting the historical content is solely based on the date range—

December 6, 1941 to December 8, 1941. Selected materials include: documents, diaries, telegrams, letters, memoranda, and photographs.

Library staff has endeavored to make this research collection as comprehensive as possible; this collection does not represent the entirety of n

nts of **Pearl Harbor**. There is a vast amount of documents relating to the lead up to **Pearl Harbor** attack itself, and the aftermath. To simpli

available within this collection. **The Pearl Harbor Guide** is available for researchers seeking additional information, including documents relat

WWII.

Provenance: The Pearl Harbor artificial collection includes: **President**'s **Official File** (OF), **President**'s Personal File (PPF), **President**'s **Secr**

Social Tags:

- Hawaii
- Film
- Attack on Pearl Harbor
- Pearl Harbor advance-knowledge conspiracy theory
- National Pearl Harbor Remembrance Day
- USS Arizona Memorial
- Battleship Row
- Pearl Harbor
- Geography of the United States

Entities:

- City
- Company
- Country
- Date
- Facility
- Industry Term
- Organization
- Person
- Position
- Province Or State
- Published Medium
- Region

Events & Facts:

- Armed Attack
- Generic Relations
- Person Career

Pearl Harbor Attack (Dec 6 – Dec 8, 1941)

Size: (.5 cu.ft.)

Dates: December 6, 1941 – December 8, 1941

Location of Repository: **Franklin D. Roosevelt** Presidential Library

Name of **Finding Aid Author**: **Ali Caron** & **Georgina Garcia**

Date of Creation: Summer 2011

Copyright Notice: The writings of **Franklin D. Roosevelt** within this collection are in the public domain. **The official writings** of **United Sta**
e public domain. The writings of **Eleanor Roosevelt** within this collection are subject to **Mrs. Roosevelt**'s literary estate. All other materials a
101, et seq.

Administrative Note: **Franklin D. Roosevelt Presidential Library** is the first presidential library and the only presidential library used by **a si**
, and photographs pertaining and relating to **President Franklin D. Roosevelt** (FDR). FDR was in office during the events of **Pearl Harbor**

Scope and Content: This artificial collection is composed of photocopies of original documents identified from the holdings of the Franklin D. Ro
specifically the events of three days: December 6, 7, and 8, 1941. The items are a collection of documents gathered from other collections four
The criterion for selecting the historical content is solely based on the date range—
December 6, 1941 to December 8, 1941. Selected materials include: documents, diaries, telegrams, letters, memoranda, and photographs.
Library staff has endeavored to make this research collection as comprehensive as possible; this collection does not represent the entirety of m
nts of **Pearl Harbor**. There is a vast amount of documents relating to the lead up to **Pearl Harbor** attack itself, and the aftermath. To simplify
available within this collection. **The Pearl Harbor Guide** is available for researchers seeking additional information, including documents relate
WWII.

Provenance: The Pearl Harbor artificial collection includes: **President**'s **Official File** (OF), **President**'s Personal File (PPF), **President**'s **Secre**
apers, Francis Biddle Papers, Charles Fahy Papers, William Hasset Papers, **Henry Morgenthau Jr.** Diaries, Frank A. Schuler Papers, John Tolar
Processing Notes: The collection was arranged, researched, and described in summer 2011, by interns **Ali Caron** and **Georgina Garcia** under
rvisory archivist Bob Clark.

Series Descriptions: The collection is organized in 2 series:

• Series I: Documents – The items selected for this series remain within December 6, 1941-December 8, 1941, date range. The items reflect th
The contents found under the container list portion are arranged by: collection title; and folder title, found in quotations.

• Series II: Still Photographs – Images comprising this series are selected from the **FDR Library**'s **General** Photograph Collection, folder: WW
ginal captions taken from the photographs themselves, along with a Library control number. Unless copyright information is stated in the image

he public domain. This means that, to the best of our knowledge, the materials may be freely used by **the researcher**. However, for copyright
ermine the limits of Fair Use as defined by sections 107 to 118 of the **copyright law** and to obtain permission from the copyright holder for fu

Arrangement: Series I is arranged alphabetically after **the president**'s papers and Series II is arranged numerically.

Container Lists:

SERIES I: DOCUMENTS

OF400: Appointments; **Hawaii**, 1941

OF4675: World War II; **General**, 1941-1942

PPF200b: Nov. 11, 1941- Jan. 6, 1942; Public Reactions

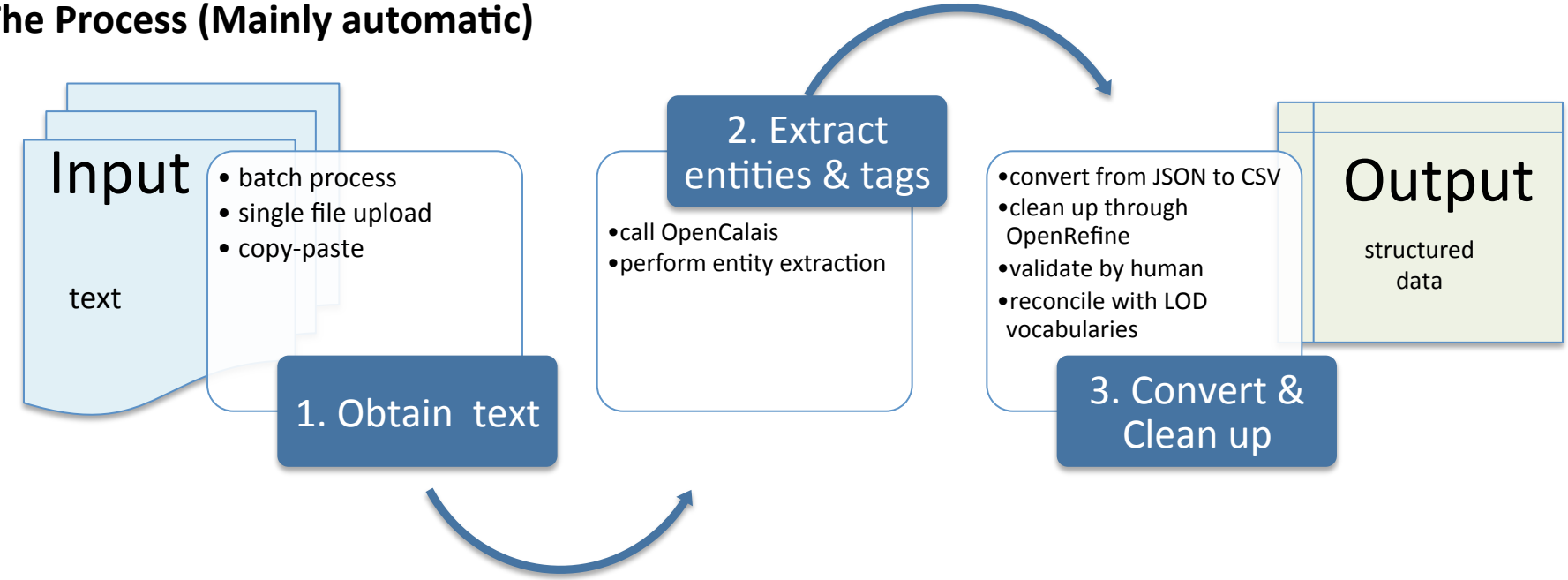
Two research samples were used to analyze the access points supplied by OpenCalais semantic analysis tool

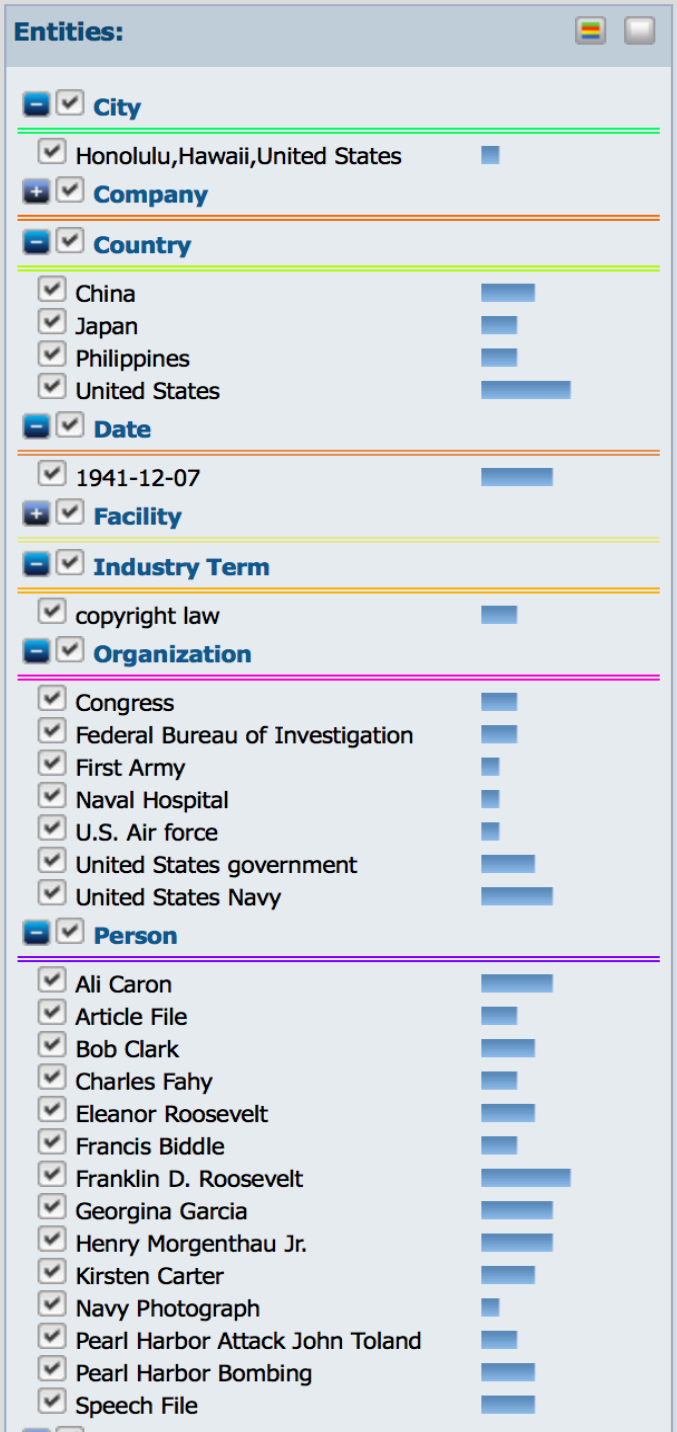
Sample 1: Finding Aids

- 43 archival record groups
- from sixteen institutions, including
  - university archives,
  - government records archives, and
  - manuscript/special collections repositories in various LAMs.

- Text from the archival finding aids
  - Descriptive information
    - creator histories
    - scope and content notes
    - detailed description of contents, including folder and item titles
  - Abstracts from these descriptions

The Process (Mainly automatic)





Original Panofsky **Object of Interpretation**

Level 1
-*Primary or natural* subject matter
(A) factual,
(B) expressional-,
constituting the world of artistic motifs

ofness

Level 2
-*Secondary* or *conventional* subject matter, constituting the world of *images*, *stories* and *allegories*.

CCO Simplified layers

1 - **Description**
(refer to the generic elements depicted in or by the work).

ofness & aboutness &
(limited from ofness aspects)

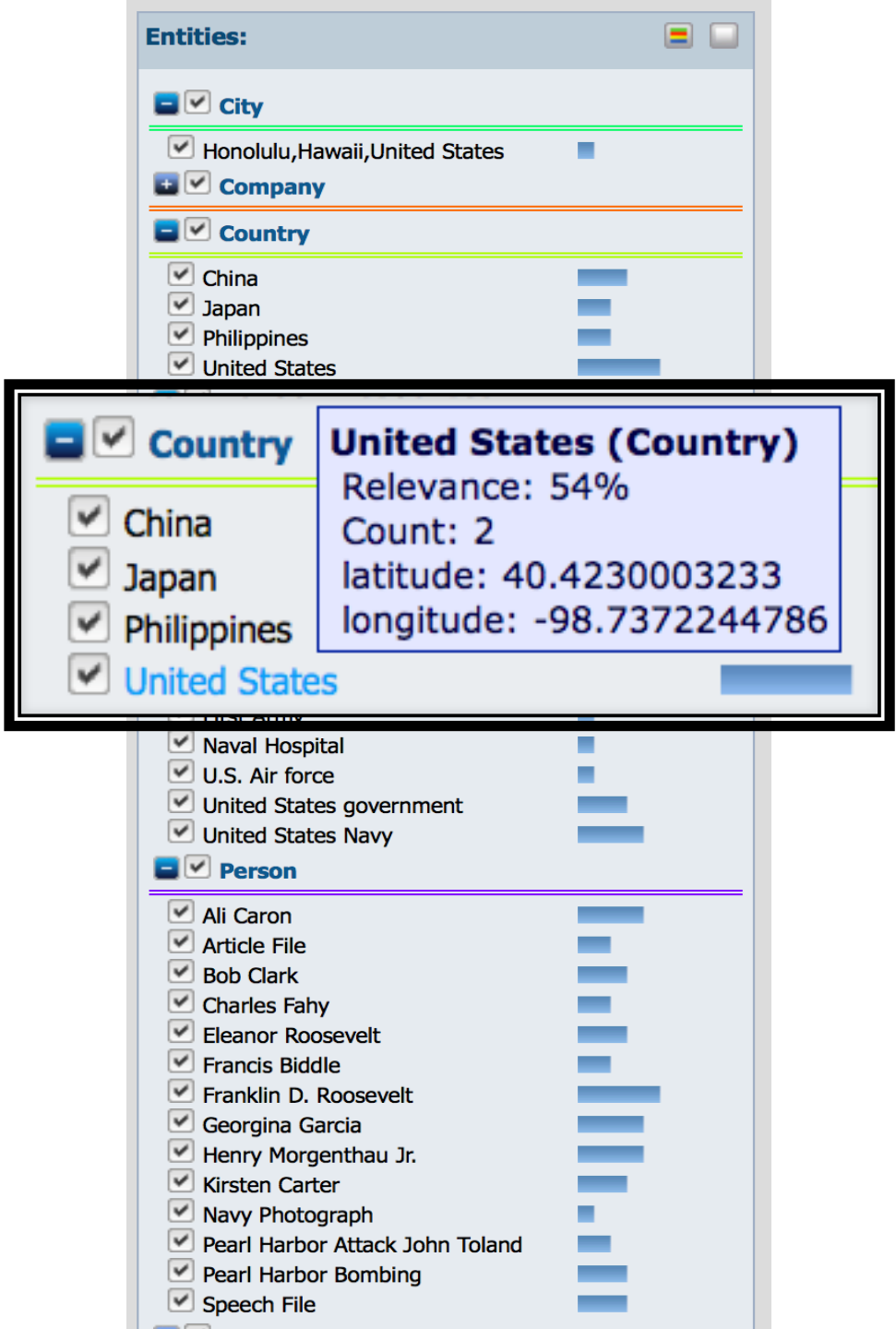
2 - **Identification**
(refer to the specific subject).

Findings

Entities correctly identified via Calais analysis (at level one, or, description) included:

- personal names (Person),
- corporate names (Company, Facility, Organization),
- geographic names (City, Continent, Country, Natural Feature, ProvinceOrState, Region), and
- events (Holiday, PoliticalEvent).

Calais provides relevance scores for each identified entity, which may be used as a valuable clue about the importance of that entity to the overall scope of the archival collection.



Findings

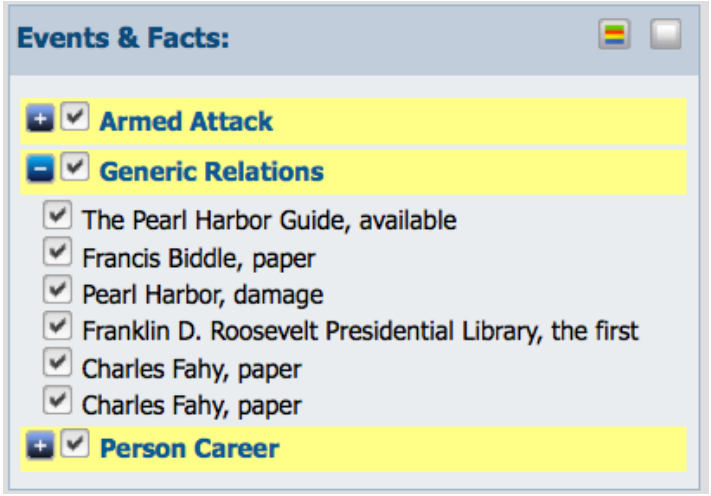
In addition to entities, Calais also generated many topical terms describing the subject matter of the records (level two, or, identification)

- These topics were often found
  - as social tags or
  - as entities under the “IndustryTerm”
  - or as entities “Product” category.

These categorizations were the least reliable in terms of accuracy:

- incorrectly identified text strings from the finding aids as *products* or *industry terms*.
- analysis of detailed description areas was most likely to lead to incorrect identification of text strings because the descriptions have the physical location information intermingled.
- Reason: the raw data (unedited) that was fed to the engine:
  - the entire finding aid was used
  - often included physical location information for the records and document formatting

Targeted analysis of particular areas of the finding aids may result in better accuracy for topical analysis.



Suggestions based on Sample 1 (Finding Aids)

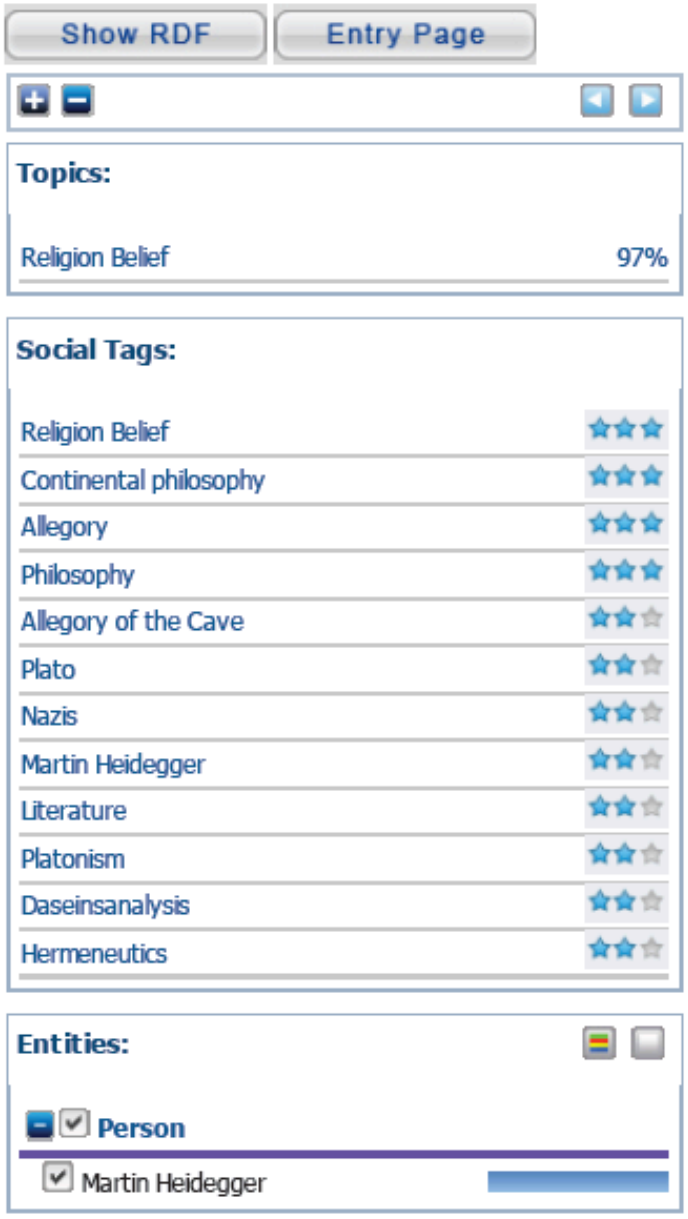
- It would be well worth the effort for institutions to experiment with semantic analysis methods as
  - an initial step to suggest key entities and topics, or
  - as a final check to ensure that important concepts or entities have not been overlooked.
- For certain types of records, particularly those for which subject indexing is not common, semantic analysis may provide entry points to archival records that were not previously available.
- Such techniques will enhance subject analysis at the first two levels (description and identification), but are unlikely to be useful for interpretation of the material.

Sample 2

- 44 philosophy theses
  - a selected sub-sample (22) from KentLINK; and
  - a random sample (22) from OhioLINK.

- Abstracts,
- titles,
- keywords, and
- introduction paragraphs

- Process (manual)
  1. Submitted to OpenCalais separately to obtain the results.
  2. All of the candidate terms were counted according to Agent Names, Geographic Names, Corporate Name, and Topic Terms.
  3. They were manually validated to determine
     1. the relevance to the thesis,
     2. the type of a term (e.g., named entity, tag, or general heading),
     3. its availability in
        1. LCNAF,
        2. LCSH,
        3. Wikipedia (as an entry), and
        4. the Stanford Encyclopedia of Philosophy.



Original Panofsky **Object of Interpretation**

Level 1
-*Primary or natural* subject matter
(A) factual,
(B) expressional-,
constituting the world of artistic motifs

ofness

Level 2
-*Secondary* or *conventional* subject matter, constituting the world of *images*, *stories* and *allegories*.

aboutness &
(very general)

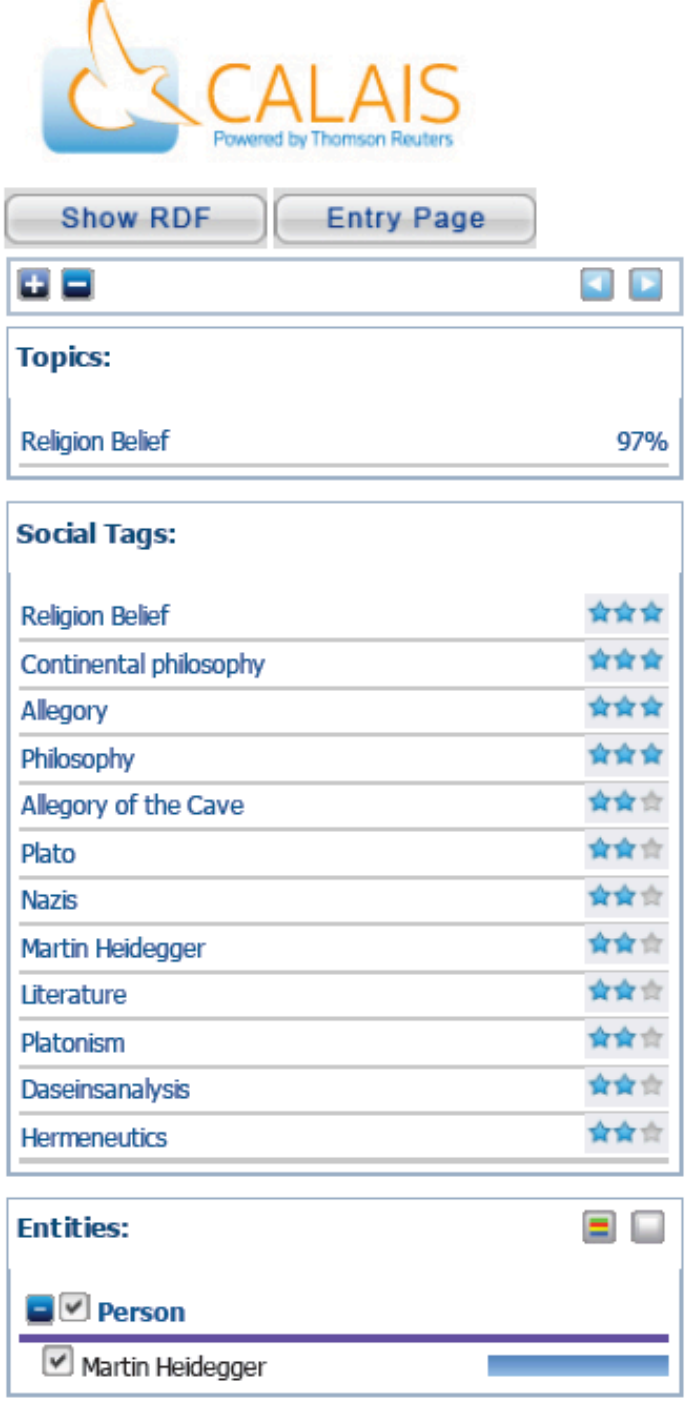
Additional level:
- Inferencing

CCO Simplified layers

1 - **Description**
(refer to the generic elements depicted in or by the work).

ofness & aboutness &
(limited from ofness aspects)

2 - **Identification**
(refer to the specific subject).



Research Findings from Sample 2

- Semantic analysis based on the abstracts generated more successful tags than those based on the titles.
- Some entity names missed in the Entity section were often correctly extracted into the tags section
  - E.g., singular names such as Plato and Aristotle, or
  - instances where the first name was not included
- Major concepts were correctly identified in most cases.

Research Findings from Sample 2 (cont.)

- The software often over-generalized the subjects by assigning very general terms (e.g., "philosophy," for almost every philosophy thesis) and
- Some terms that were unrelated to the subject of the thesis.

KentLINK sub-sample:
- average 9 tags per abstract,
- an average of 1.64 were overly broad topical terms and
- an average of 3.45 were unrelated topical terms (slightly more than 1/3).

OhioLINK sub-sample had similar figures.

- This level is different from "identification" and "description", seems to be "inferencing".

Suggestions based on Sample 2

- Level I "description" -- the tags did very well
- Level II "identification" – adequate
- The tags that could be categorized as "inferencing" results seemed to be less valid according to the best practices of cataloging and subject indexing.
  - The overly-broad topic terms are not wrong (e.g., *philosophy*, *knowledge*, *science)* but their relevance in terms of subject access is questionable.
- The promising news: among the topical terms (including named entities as topics),
  - LCSH together with LCNAF could match about 75% of them closely (we used the degree as closeMatch, in comparison to broadMatch, narrowMatch or noMatch),
  - DBpedia matches almost 98% with closeMatch degree for both sub-samples.
- These vocabulary sources hold great potential for these subject access points to become the linking point to the Linked Data datasets that use DBpedia and LC vocabulary URIs as their basis.

Future Research

Need User Studies

Most helpful, good, high exhaustivity	adequate, (depending on the domain and raw data)	Not applicable	maybe useful
I. description (ofness)			
	II. identification (aboutness & ofness)		
		III. interpretation (aboutness)	
			inferencing (aboutness)

Acknowledgement

- The authors would like to thank research assistants Sammy Davidson and Laurence Skirvin of Kent State University for assisting with OpenCalais-related processes.
- The research is a sub-project of Metadata-Junctions, a project funded by IMLS National Leadership Grant, 2011-2013.